

MAS2317/3317: Problems sheet 1

Introduction to Bayesian Statistics

PROBLEMS SHEET 1 is to be used in this week's problems class (4pm, Thurs 13th Feb). Further help will be available in the drop-in session the week after (4pm, Thurs 20th Feb).

Answers to the starred (*) questions should be placed in the homework submission box in the foyer of the Maths & Stats General Office (Herschel Building 3rd floor) by no later than **4pm, Tuesday 25th February**. Questions marked with a P are to be used in student presentations.

- 1.* (a) Mark on a 0 – 1 probability scale the approximate positions of the following events. Undertake a simple investigation if necessary. Explain the values you choose and state, in each case, whether you are using the frequency, classical or subjective interpretation of probability.
 - (i) A student taking MAS2317/3317 has a body mass index (BMI) of at least 22.5.
 - (ii) Newcastle United will finish in a top 6 position in the Premier League this year.
 - (iii) A randomly selected number between 10 and 99 (inclusive) is divisible by 7.(b) For each event in part (a), give a range which you believe contains the 'true' value with probability 0.95.
- 2.^P An insurance company classifies car drivers as X , Y or Z . Experience indicates that the probability that a class X driver has at least one accident in any one year is 0.02, whilst the corresponding probabilities for class Y and class Z drivers are 0.04 and 0.10 respectively. They have also found that, of the drivers who apply to them for cover, 30% are class X , 60% are class Y and 10% are class Z .
 - (i) Mr Smith is a new client and, within 12 months, he has an accident. What is the probability that he is a class Z risk?
 - (ii) If Mrs Brown goes for n years without an accident, and we assume the number of accidents in different years to be independent, how large must n be before the company considers that she is more likely to belong to class X than to class Y ?
- 3.^P Suppose that 1% of the population have a disease D . A diagnostic test S , designed to detect the disease, has the following accuracy:

$$\begin{aligned}\Pr(S \text{ positive} | D) &= 0.95 & \text{and} \\ \Pr(S \text{ positive} | D^c) &= 0.05.\end{aligned}$$

If 100 people were tested at random, how many would we expect to test positive and of those that do, about how many would we expect to have the disease?

4. An experiment can result in four possible outcomes x_1 , x_2 , x_3 and x_4 . The probability of each of these outcomes is affected by a parameter θ which can take one of three values θ_1 , θ_2 and θ_3 :

$p(x \theta)$	x_1	x_2	x_3	x_4
θ_1	0.1	0.2	0.3	0.4
θ_2	0.3	0.3	0.2	0.2
θ_3	0.5	0.1	0.1	0.3

The prior plausibilities attaching to θ_1 , θ_2 and θ_3 are 0.1, 0.4 and 0.5. Find the posterior probability function for θ after observing x_1 . Repeat the calculations for x_2 , x_3 and x_4 .

- 5.* On June 1st 2009, Air France flight 447 disappeared en-route to Paris Charles de Gaulle Airport about four hours after leaving Rio de Janeiro. Locating the aircraft's flight data recorder (the *black box*) was key to understanding the cause of one of the deadliest accidents in the history of Air France.

After considering various oceanographic characteristics in the Atlantic at the time the flight disappeared, initial searches for the data recorder focussed on a large area to the north of the point at which it was believed the aircraft made impact with the ocean. In fact, it was believed that there was an 80% chance of finding the data recorder in this area, as opposed to the south of the impact zone. After almost two years of searching in the area mainly to the north of the point of impact, the elusive black box had still not been found.

In April 2011 *Metron, Inc.* were hired to launch a Bayesian review of the search effort. The following probabilities of black box “identifiers” were estimated, based on the black box being in the areas defined as “North” and “South”:

	Identifier			
	x_1	x_2	x_3	x_4
North	0.3	0.1	0.1	0.5
South	0.05	0.7	0.2	0.05

These “identifiers” were:

- x_1 : “Fresh” debris found (since December 2010)
- x_2 : Signal detected from the black boxes
- x_3 : Both “fresh” debris found *and* signal from black boxes detected
- x_4 : Neither “fresh” debris found *nor* signals detected from the black boxes

- (a) Construct rules for differential diagnosis of the most plausible search area for the flight's data recorder, on the basis of the posterior assessments for each type of identifier.
- (b) According to your set of rules in part (a), what is the probability of ‘misdiagnosis’?
- (c) In March 2011 a weak signal from the black box was detected. Quantify the shift in odds in favour of searching in the area south of the impact zone, having observed this identifier.

6. Extensive manuscripts of two scribes show that they differ on the relative frequencies with which they use two alternative vowel forms e and oe , as in the following table:

	e	oe
Scribe A	0.3	0.7
Scribe B	0.5	0.5

A new manuscript fragment, known to be by one of the scribes, uses the vowel form five times, three times as e and twice as oe . If an independent assessment of similar fragments favours Scribe A as author with probability 0.7, show that the evidence from the vowel form is sufficient to turn the odds in favour of Scribe B.

7. If $X_1, X_2, \dots, X_n$ are independent random variables with probability function

$$f(x|\theta) = \theta(1 - \theta)^{x-1}, \quad x = 1, 2, \dots,$$

show that $T = \sum_{i=1}^n X_i$ is sufficient for θ .

8. If $X_1, X_2, \dots, X_n$ are independent $N(5, \sigma^2)$ random variables, show that

$$T = \sum_{i=1}^n (X_i - 5)^2$$

is sufficient for σ .

- 9.* Hurricane-strength wind speeds (X miles per hour) for locations in the Gulf of Mexico are modelled using a distribution with probability density function

$$f(x|\sigma) = \frac{x}{\sigma^2} \exp \left\{ -\frac{x^2}{2\sigma^2} \right\}, \quad x > 0, \quad \sigma > 0.$$

During a hurricane, you collect wind speeds at n randomly chosen locations in the Gulf of Mexico, giving $x_1, x_2, \dots, x_n$. Find the likelihood function for σ , and obtain a sufficient statistic for σ .